研究プロジェクト報告

史的文字データベース連携検索システムの公開

井上　聡

はじめに

本誌八八号にて紹介した、歴史的字形を対象とする多機関連携検索が、三月末に実証試験版として公開の運びとなった。当初、国内機関のみならず台湾の中央研究院歴史語言研究所等が擁するコンテンツも加えて、国際的な連携ポータルサイトを開設することを目指していた。あわせて本連携のコンセプトについてマスコミを介して広く発信することなども企図していたが、新型コロナウイルスの流行に伴う状況に鑑み、まずは国内連携に限定する形で、試験的なサイトの開設に踏み切った次第である。国際的な連携構築や、広汎な参画の呼びかけといったアクションについては、社会情勢の安定を踏まえて、今後可及的速やかに取り組んでゆきたい。本号にては、まずこのたび公開したサイトのあらましを紹介し、現状と課題を述べておきたい。

一　連携検索の対象データについて

今回の連携検索用ポータルサイトの開設にあたっては、奈良文化財研究所が幹事となって推進し、史料編纂所・国文学研究資料館のほか、国立国語研究所・京都大学人文科学研究所・中央研究院歴史語言研究所／数位文化中心が協力する体制をとった。現状で検索可能なデータ群は、木簡庫（奈良文化財研究所）・電子くずし字字典（史料編纂所）・日本古典籍くずし字データセット（国文学研究資料館）が有するコンテンツとなる。それぞれが擁する歴史的字形の画像数は、木簡庫が約一〇・五万件（文字種約一八〇〇）、電子くずし字字典が約二九万件（文字・語彙種約八五〇〇）、日本古典籍くずし字データセットが約一〇九万件（文字種約四三〇〇）を数え、総計一五〇万件弱に達する。平城京の木簡から近世版本にいたるまで、奈良時代から江戸時代におよぶ千年余の字形を網羅している。今後、国際的連携も進めることで、東アジア漢字文化圏を網羅する連携へと発展させてゆくことを目指している。

二　連携検索用ポータルサイトの概要

連携検索用ポータルは、幹事部局である奈良文化財研究所システム内に構築されており（https://mojiportal.nabunken.go.jp/ja）、史料編纂所ＨＰからもリンクしている。今後、連携に参画する各機関にて、それぞれの目的に応じた独自のポータルが作られてゆくことになるだろう。

図１　検索画面

今回のサイト画面設計は、従前より奈良文化財研究所と史料編纂所の間でなされてきた連携検索画面を基本としたシンプルな構造になっている。検索画面は、調べたい文字を１字入力する形をとり（**図１**）、サイトからAPIを介して各機関のデータ群に照会する設計となっている。検索結果は、機関ごとの回答を左右方向に一列にならべ、全体を一画面に集約して表示している（**図２**）。

従前の連携検索と決定的に異なる点は、各機関が字形画像データをIIIF形式に転換することを前提として、設計がなされていることである。検索結果として字形画像のみならず、当該データのメタデータを記述したマニュフェストファイルが

図３　１件ごとの表示

図２　検索結果画面

返却されることで、ユーザーは得られた結果を多様な視点から分析することが可能になる。以下、検索結果画面からの遷移をたどることで、具体的に説明を加えておきたい。

各機関より回答された字形データは、１件ごとに見ると**図３**のような形で表示される。画像の下には、３つのボタンが付されており、それぞれ異なった役割を担っている。まず左端の「詳細」とあるボタンであるが、これは典拠となった各データベース画面へのリンクである。オリジナルとなるデータベースに遷移することで、当該データにどのような情報が与えられているか詳細を確認することができる。これは従前の連携検索からの継承である。

対して残る２つのボタンは、IIIF を導入したことで新たに付与された機能である。中央のアイコンを押下すると、字形画像ごとに付されたマニュフェストファイルが表示される（**図４**）。各種メタデータが、IIIF 仕様に準拠した形で記述されており、これを活用することで、ユーザーは字形画像を、任意の IIIF 対応のビューアに取り込んで自在に二次利用することが可能になる。

このマニュフェストをもとに対応ビューアの一つである Mirador にて表示するのが右端のアイコンである。押下すると**図５**のような画面が展開し（図は木簡庫の例）、当該字形の抽出元画像とともに、切り出された字形の範囲が明示される。右側には対象資料や文字に関するメタデータ・注記のほ

図４　マニュフェストファイルの表示

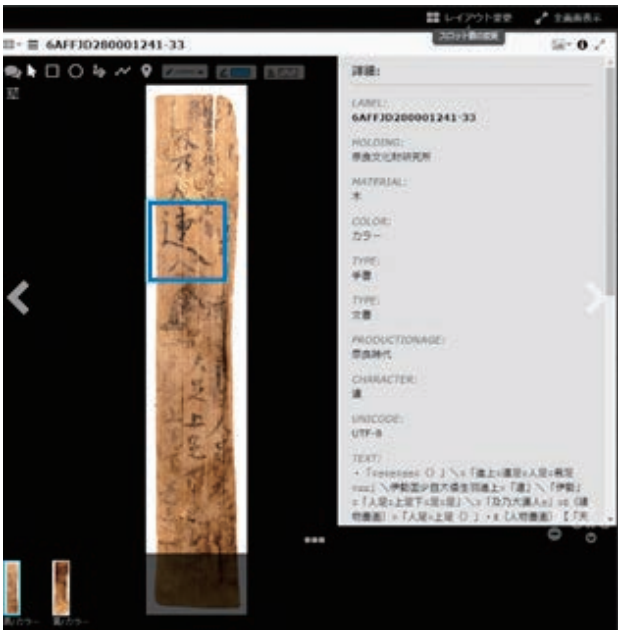

図５　Mirador による結果表示

か、データの利用条件などが表示される仕組になっている。さまざまな典拠データベースから抽出されたデータが、マニュフェストを介して共通のフレームのうちに表現されることで、より丁寧な分析が可能になると言ってよいだろう。

三　字形データの応用的展開と今後

今回の連携では、データのオープンリソース化とIIIFの導入を基軸としてサイト構築を実践したが、これによって生じる変化とは何なのだろうか。検索対象データの拡充も大きなポイントになるが、焦点はやはり先にも記したマニュフェストファイルの活用になるだろう。

IIIF形式で公開されているものであれば何であれ、連携検索から導かれた字形画像と並べて比較することが可能になる。例えば、**図6**は国立国会図書館のデジタルコレクション（同館所蔵「敦煌等經文」）と木簡庫のコンテンツを並べてみた。必要な手続きは、Mirador上に、それぞれのマニュフェストをドラッグ&ドロップするだけである。この事例では「無」という字形の比較を意図している。ある機関のサイトにて史資料を見ている際に、こうした比較・対照がすぐさまできるとすれば、研究は大いに効率化されるだろう。

今後、本連携にあっては、国際的にも拡大を図り、検索対象となる字形の質・量をさらに高めてゆくこと、さらに単文字検索にとどまらない検索条件の拡張に努めてゆくことなどが課題となってくる。また典拠となる各データベースにおいても、オープン化に根差した弛みない改善が必須となる。

今回開設したサイトは、まだ試験版に止まるもので、改善の余地が大きく残されていることは、関係者一同が強く認識している。読者各位にあっては、本サイトを様々にご活用いただき、お気づきの点・ご要望などについて、忌憚ないご意見をいただければ幸いである。

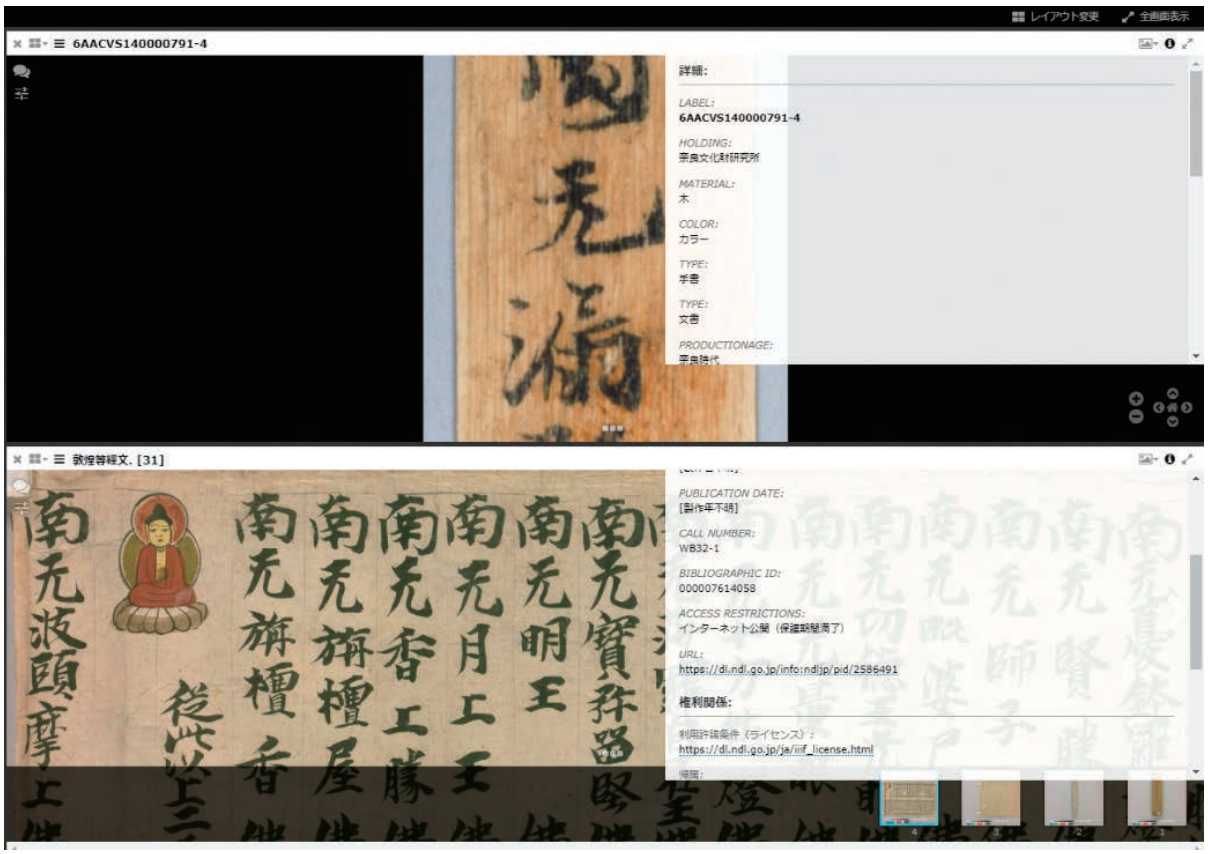

図6　木簡庫と国会図書館所蔵史料の同一画面表示